VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

Virginia Bioinformatics Institute
Bioinformatics Resource Center
http://brc.vbi.vt.edu

October, 2004

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

Virginia Bioinformatics Institute

- Created in July, 2000
- Part of Virginia Tech (Blacksburg, VA)
- 16 faculty members and approx 200 research scientists and staff
- Projects related to BRC
  - Pathport project (DoD)
  - MARCE (bioinformatics and genomics research core)
  - proteomics resource database

Our organisms

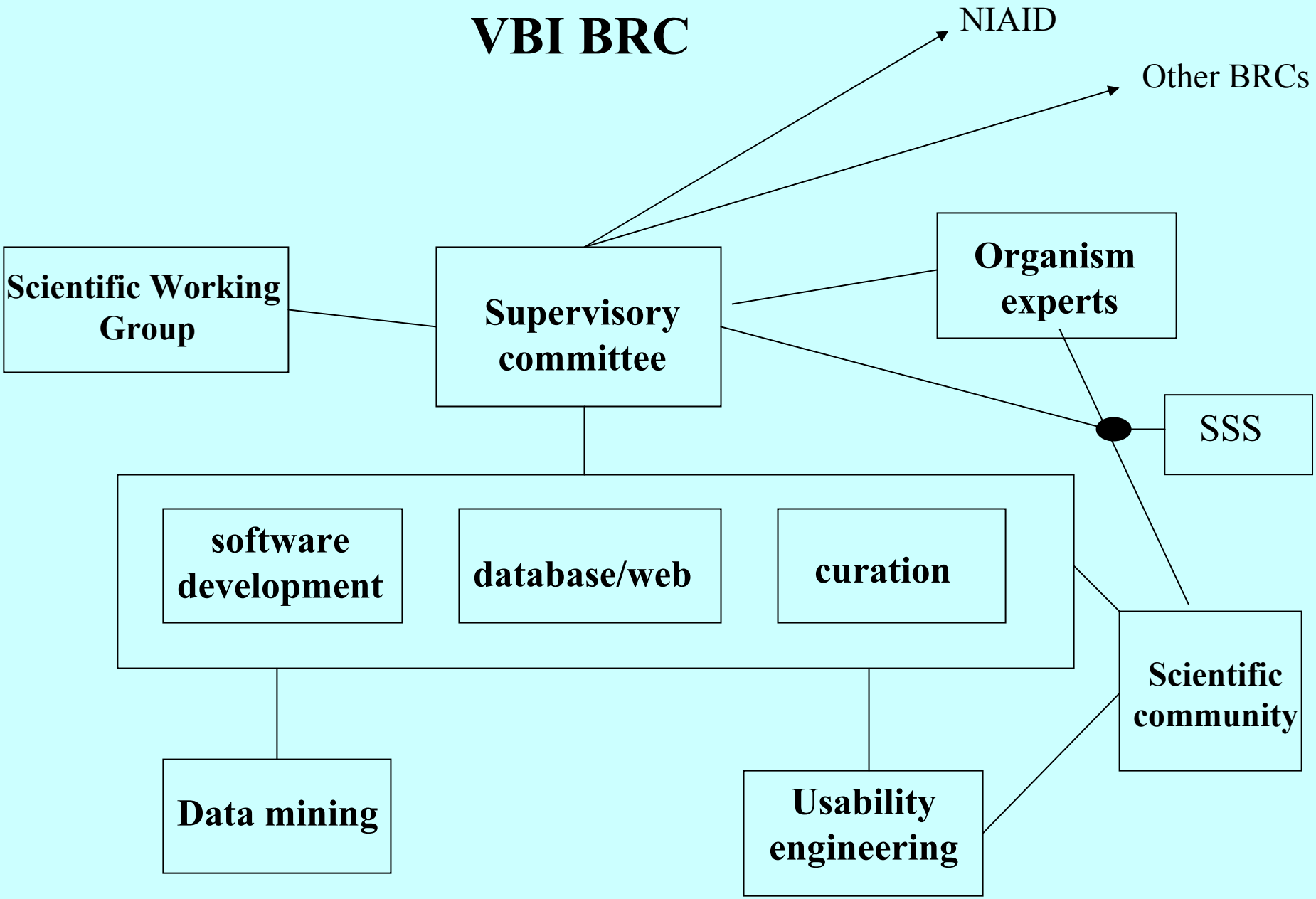
- **Bacteria**
  - *Rickettsia*
  - *Brucella*
  - *Coxiella burnetii*
- **Viruses**
  - Coronaviruses (SARS)
  - Caliciviruses
  - Rabies viruses
  - Hepatitis A viruses
  - Hepatitis E viruses

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

Current count: **238 genomes**

	viruses	**bacteria**
Number of genomes	95%	5%
Number of genes (approx)	20%	80%

VBI BRC


NIAID
Other BRCs
Scientific Working Group
Supervisory committee
Organism experts
SSS
software development
database/web
curation
Data mining
Usability engineering
Scientific community

Organizational structure

- **Supervisory committee**: Bruno Sobral (PI), João Setubal (coPI), Project Manager, group leaders
- **Curation group**: Oswald Crasta
- **Database/Web group**: J. Setubal
- **Software development group**: Dana Eckart
- **Collaborators**: Debby Hix and Joe Gabbard (SRC/ VTech, usability engineering) and Naren Ramakrishnan (CS/VT, data mining)
- **Scientific Working Group:** 10 external scientists, advisory role, not yet created

Organism experts

organism	**expert**	**affiliation**
Brucella	Stephen Boyle	VTech
Rickettsia and *Coxiella*	Abdu Azad	U Maryland
coronaviruses	Susan Baker	Loyola
Hep E viruses	Xian-jin Meng	VTech
Hep A viruses	Greg Armstrong*	CDC
Rabies virus	Charles Ruprecht*	CDC
Calicivirus	Stephen Monroe*	CDC

* awaiting confirmation

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

BRC Mission

- Provide curated data
- Provide analysis resource
- Facilitate the discovery of new ways to fight infectious diseases

Curation

- **Annotation**: what is "traditionally" done for genomes; what you get when you download bacterial/viral genome files from NCBI
- **Curation**:
  - Make annotation as accurate and as standard as possible
  - Enrich with additional sources of information
    - Literature
    - Other experimental data ("post-sequence data")
    - Input from experts in the scientific community

Our general approach to curation

- We will use the concept of ***reference genome*** (RG) and ***associated genome*** (AG)
- For **each class**, we'll choose **one genome** and carefully curate it: **RG**
- Each other genome in the class will be declared an AG, whose curation will be derived from the RG
- Choice of RG to be made with input from scientific community

Four steps to RG curation

1. Genome annotation: prediction/detection of various features (protein-coding genes, rRNA genes, repeats, etc)
2. Gene annotation
   - In-house automated pipeline
   - Comparison between pipeline results and existing annotation
   - Comply with standards
   - Create a "UniGene set" for its class
3. Literature-based gene curation
4. Integration of post-sequence data

Associated Genome curation

- "import" annotation from RG (requires accurate ortholog sets)
- Determine *unique* genes, features
- Enlarge UniGene set for the class with unique genes
- When all AGs have been curated, major release of curated genomes for the class
- Selected AGs/gene groups may undergo special curation (as requested by the community)

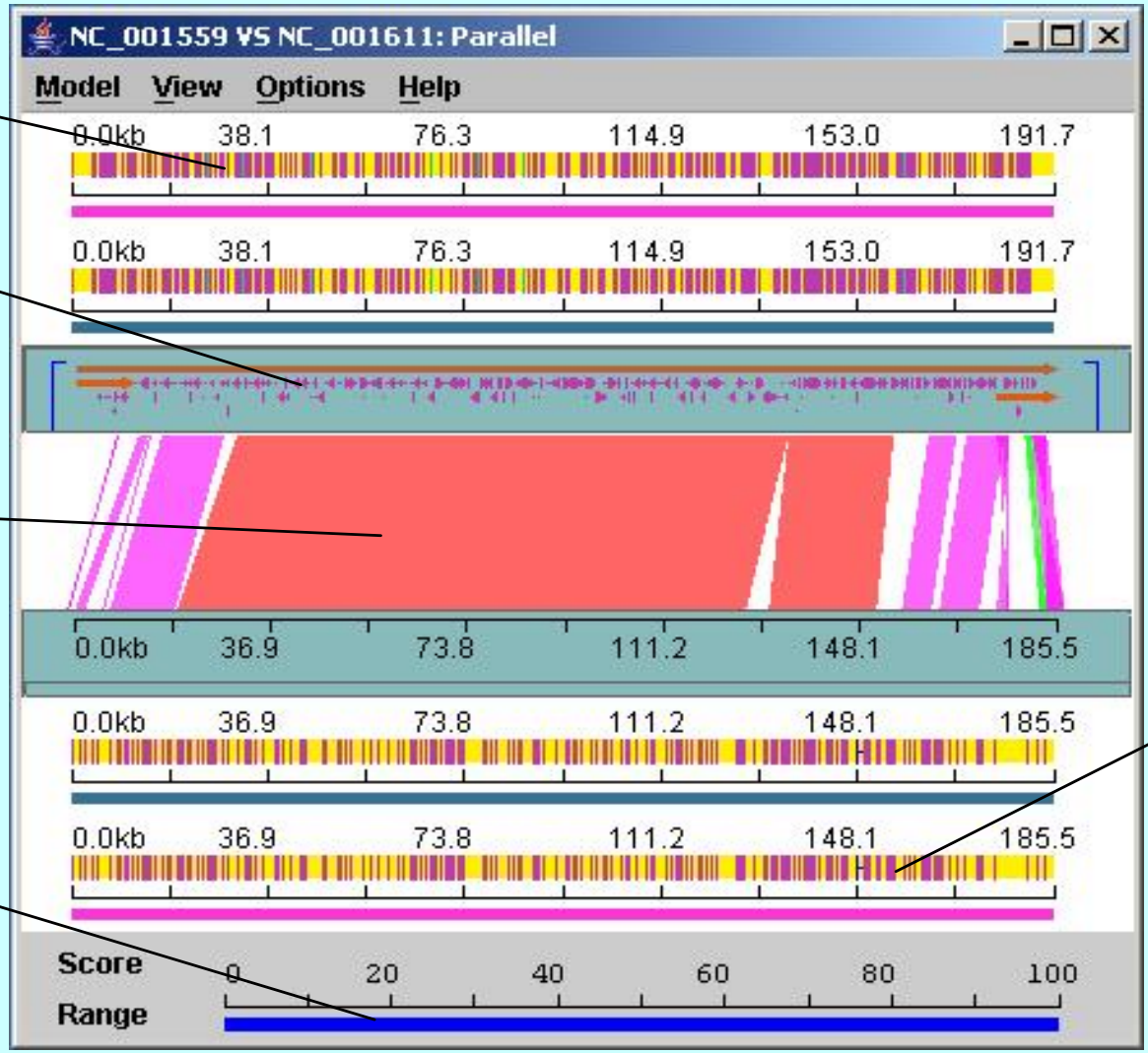
Comparative genomics

Three levels

1. Whole genome (e.g. alignments)
2. Gene-centric (e.g. ortholog sets, unique genes)
3. Sequence variation (codon usage differences, SNPs, etc)

Analysis resource

- Web-based, fully cross-linked
  - Genome and gene families/grouping browsing
  - Sequence search
  - Database query
  - Comparative genomics pages
  - Easy access to post-sequence data
- toolBus resource
  - Client-side interconnect; gives access to and integrates many 3rd party genome analysis tools
  - Developed at VBI; already functional
  - provides capabilities similar to those planned for the web-based interface



Sequence 1

Features

Matching region

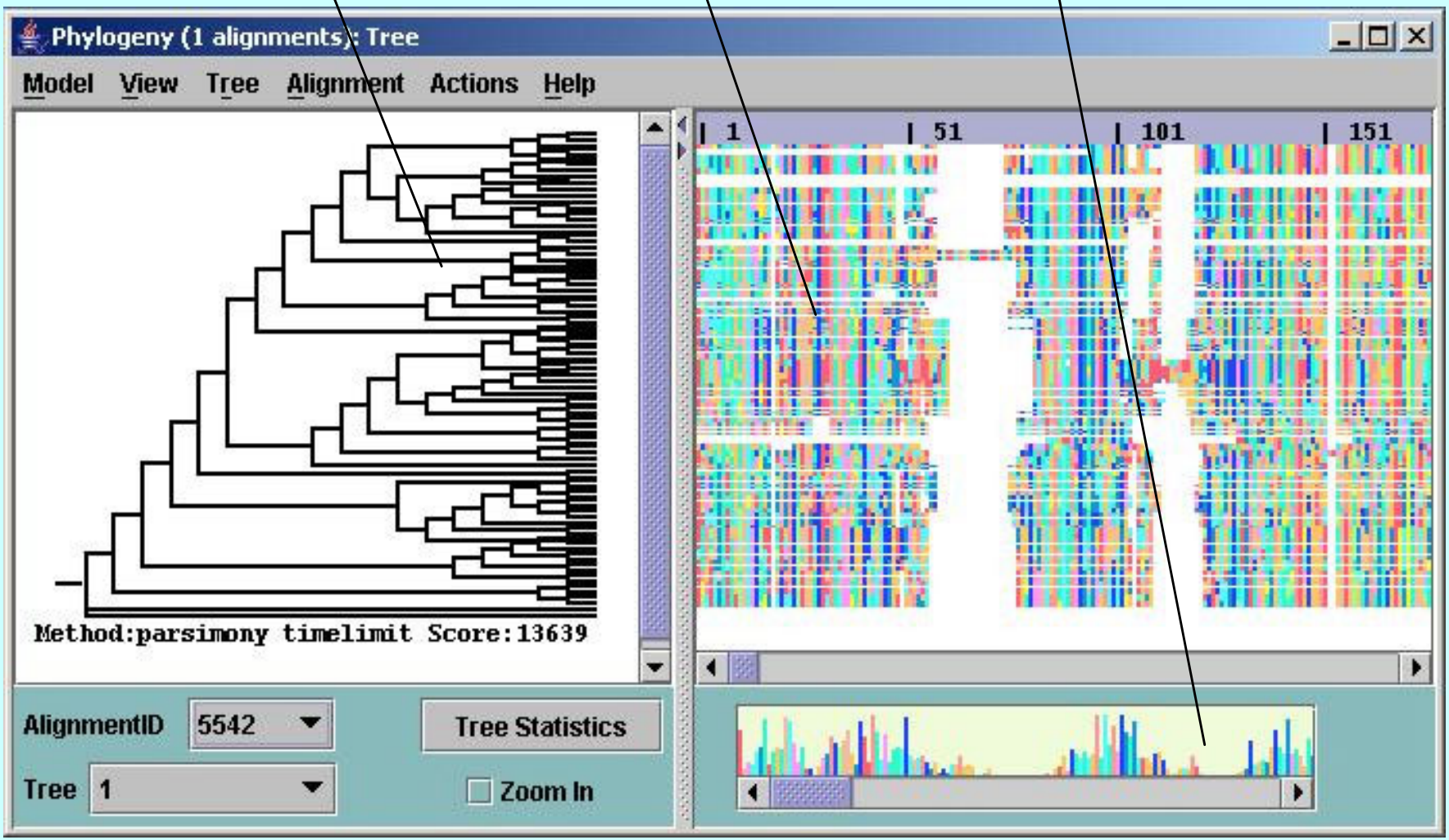
Scoring range selector

Sequence 2

Phylogenetic tree

Sequence alignments

Sequence region selector



VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

General timeline for first 2 years

1. **Dec 04**: **AsIs project**: make existing annotations available through website
2. **Jun 05**: Genome/gene annotation for all RGs
3. **Jun 06**: AGs curation and literature-based curation (first release, ongoing with refinements and additional genomes)
4. **following years**: Post-sequence data curation

Interactions with scientific community

- *Organism experts*
  - Active contributors in the curation process
  - Will select genes of special importance
  - Will help with literature search
  - Bridge to specific scientific communities
- *Genome providers*
- *General scientific community*
  - A good protocol is being currently worked on

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

Database architecture

- DBMS + schema
  - Oracle + GUS (Genomics Unified Schema, developed at UPenn)
- Why GUS?
  - Usable and available now; meets requirements
  - At least 3 other groups at VBI are using it (considerable local expertise)
- Curation database + public database

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

Some VBI BRC principles

- Use as many existing analysis/viewing tools as possible (example: GBrowse, bioperl, etc)
- In terms of curated data we want to be comprehensive and accurate…
- …however, we have to balance manual curation (high accuracy) with automated analysis (high throughput, hence scalable)
- User interfaces: keep them simple and useful (usability engineering)

Size of the VBI BRC operation

- In addition to group leaders (1 FTE), project manager (1 full-time), and external collaborators (0.5 FTE):
  - 4 full time curators
  - 1 full time database analyst
  - 4 full time software developers

Issues under study

- What is the depth of curation?
- Differences between viral and bacterial curation
- How best to benefit from interactions with the scientific community
- Interaction with sister BRCs (see interoperability session later)

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

Conclusion

- We look forward to our 6th year of operation!